TESINA PARA OPTAR POR EL GRADO DE
LICENCIADO EN CIENCIAS BIOLÓGICAS

GENERACIÓN DE UNA BASE DE DATOS COMPLETA Y NO REDUNDANTE DE VIRUS Y SU ANÁLISIS COMPOSICIONAL

Diego Simón

dsimon@fcien.edu.uy

Orientador:

Dr. Héctor Musto

Laboratorio de Organización y Evolución del Genoma
Departamento de Ecología y Evolución
Instituto de Biología
Facultad de Ciencias
Universidad de la República

Tribunal:

Dr. Juan Arbiza
Dr. Andrés Iriarte
Dr. Héctor Musto

Diciembre de 2015

Resumen

Las firmas genómicas no dependen de regiones codificantes y son características del genoma en su conjunto. Los mecanismos aún no están entendidos completamente. Los virus están expuestos a sesgos mutacionales propios del hospedero. Aquellos más expuestos a estos sesgos, presentarán efectos más evidentes. El objetivo de esta tesina es explorar la diversidad viral e intentar describir grandes patrones evolutivos. Utilizando un abordaje bioinformático, se trabajó con 4195 genomas completos de *Viral Genomes*, del NCBI. Se calcularon composiciones de bases y frecuencias de dinucleótidos, por virus y por tipo de genoma, estructura, grupo de Baltimore, orden, familia y hospedero. En muchas de las categorías se observaron empobrecimientos en TpA. El grupo *VI* y los órdenes *Ligamenvirales* y *Nidovirales* no presentaron sesgo para TpA. La única familia con enriquecimiento en TpA fue *Globuloviridae*, familia de virus de arqueas termófilas. Sesgos en CpG fueron observados en virus de ARN, pero no en el conjunto de todos los virus de ADN, probablemente como consecuencia de la gran cantidad de bacteriófagos. Los grupos *IV*, *V* y *VI*, los tres de ARN de hebra simple, presentaron empobrecimiento en CpG. Virus de animales y de plantas también. Además de su gran variabilidad, deben considerarse otros factores como el tamaño de los genomas o la presencia de actividad *proofreading*. Aspectos que dependen del hospedero, como los sesgos mutacionales, también son determinantes.

Palabras clave: virus, análisis composicional, frecuencia de dinucleótidos

Abstract

Genomic signatures do not depend on coding regions and are characteristic of the whole genome. The mechanisms are not yet fully understood. Viruses are exposed to the mutational biases of their hosts. Those most exposed to these biases, will present more noticeable effects. The aim of this work is to explore the viral diversity and try to describe major evolutionary patterns. Using a bioinformatics approach, we worked with 4195 complete genomes from Viral Genomes database. Base compositions and dinucleotide frequencies for viruses and by genome type, structure, Baltimore´s group, order, family and host were calculated. In many categories underrepresentations in TpA were observed. Group VI and orders *Ligamenvirales* and *Nidovirales* showed no TpA bias. The only family with TpA enrichment was *Globuloviridae*, a family of thermophilic archaeal viruses. CpG biases were observed in RNA viruses, but not in the set of the entire DNA group, probably a consequence of the large number of phages. Groups IV, V and VI, the three ssRNA groups, showed CpG underrepresentation. Also animal viruses and plant viruses. Besides its great variability, other factors should be considered such as the size of the genomes or the presence of proofreading activity. Host dependent aspects like mutational biases are also crucial.

Keywords: viruses, compositional analysis, dinucleotide frequency

INTRODUCCIÓN

Los virus son entidades infecciosas generalmente de muy pequeño tamaño. Tienen como característica fundamental que solamente pueden reproducirse dentro de una célula hospedera; infectan células vivas y utilizan la maquinaria de síntesis de dichas células. Fuera de las células, los virus existen como partículas llamadas viriones, que consisten de por lo menos un genoma de ADN o ARN rodeado de una envoltura proteica llamada cápside.

Estudios de diverso tipo (particularmente metagenómicos) muestran que son las entidades biológicas más abundantes del planeta [**1**]. Infectan a todos los seres vivos (i.e., arqueas, bacterias y eucariotas). Incluso parasitan a otros virus. David Baltimore agrupó en seis grupos a las diferentes familias de virus en función de su tipo de genoma y su estrategia replicativa [**2**]. La clasificación actual comprende siete grupos (*I* al *VII*) [**3**].

El *International Committee on Taxonomy of Viruses* (ICTV) se encarga de establecer la taxonomía de los virus. Esta taxonomía se organiza en órdenes, familias, subfamilias, géneros y especies. Existen actualmente siete órdenes: *Caudovirales, Herpesvirales, Ligamenvirales, Mononegavirales, Nidovirales, Picornavirales* y *Tymovirales*. ICTV, en su última versión, incluye solamente 26 de las 104 familias en alguno de los órdenes, quedando sin asignar las restantes 78 (ver **Cuadro 1**) [**4**].

Cuadro 1. Órdenes y número de familias clasificadas por el *International Committee on Taxonomy of Viruses* (ICTV) en su última versión de julio de 2014 [**4**].

Orden	Nº de familias
Caudovirales	3
Herpesvirales	3
Ligamenvirales	2
Mononegavirales	5
Nidovirales	4
Picornavirales	5
Tymovirales	4
sin asignar	78

Existen firmas genómicas basadas en frecuencias de dinucleótidos que no dependen de sus regiones codificantes; son características del genoma en su conjunto. Los mecanismos que determinan estas firmas aún no están entendidos completamente. Pueden estar involucrados múltiples mecanismos moleculares que involucren replicación, reparación y procesos de modificación del ADN (como la metilación). También se invocan propiedades estructurales como energías de apilamiento y tendencias conformacionales del ADN. Las firmas genómicas son altamente invariables en todo un genoma y son similares entre especies cercanas [**5**].

Los dinucleótidos pueden presentar sesgos. El dinucleótido TpA se encuentra sub-representado en eucariotas y procariotas, con excepción de algunas arqueas, genomas mitocondriales de metazoarios, hongos y plantas y genomas de cloroplastos. Este sesgo se ha relacionado a la baja energía de apilamiento de TpA, la menor de todos los dinucleótidos. TpA sería contra-seleccionado debido a que tenderían a desestabilizar la doble hélice de ADN [**6**].

En los vertebrados, existe un sesgo mutacional en donde CpG es blanco de metilaciones en la C, que en caso de ser desaminada se transforma en TpG (y por complementariedad de bases en CpA, tras una ronda de replicación) (ver **Figura 1**). También en aves, aunque menos estudiadas que los mamíferos, se han descrito patrones análogos de metilación [**7**]. Estos sesgos también han sido observados en plantas [**8**].

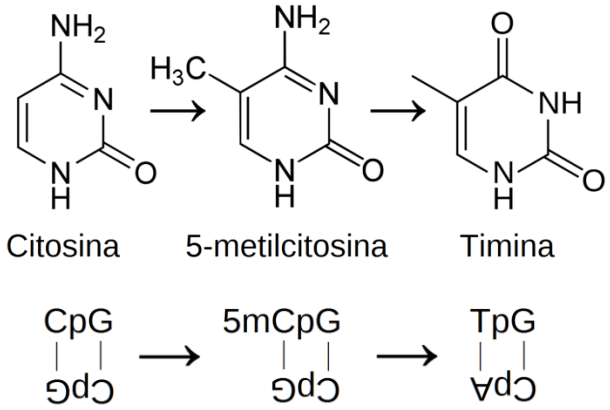

Figura 1. Esquema del cambio químico y de secuencia que provoca la desaminación de un CpG metilado en genomas de vertebrados.

Sin embargo, la metilación de C en CpG no permite explicar la sub-representación de CpG observada en los genomas mitocondriales de todos los animales. Los invertebrados, que no poseen actividad metil-transferasa y metilasa y no exhiben patrones de metilación en sus genomas nucleares, sí presentan sesgo de CpG en sus genomas mitocondriales [**6**].

El sesgo en CpG también ha sido relacionado con la energía de apilamiento. CpG, al contrario de TpA, tiene la mayor energía de apilamiento de los dinucleótidos; reduciendo la frecuencia de CpG se facilitaría la replicación y la transcripción [**6**].

Estos sesgos se presentan también en los virus. TpA es considerado universal; UpA en virus con genoma de ARN. Sesgos en CpG se observan en prácticamente todos los virus de pequeño tamaño de vertebrados (genomas de hasta 30 kb de longitud) [**9**]. También se ha descrito para algunos virus de plantas [**10**].

Los virus proponen algunos desafíos a los mecanismos propuestos. En el ARN, en doble hebra, la energía de apilamiento de CpG se encuentra en el medio de entre todas las energías para los dinucleótidos [**10**].

Con las tecnologías de secuenciación de “nueva” generación, disponer de gran cantidad de datos genómicos es una realidad. La bioinformática permite trabajar a gran escala y con un gran número de organismos, siendo las limitantes el hardware que se dispone y la fidelidad de las bases de datos. Una alta proporción de genomas sin clasificar lleva a que sea difícil la interpretación de los mismos y genera una gran decisión a priori: si trabajar con los que se sabe que están bien clasificados y anotados o si se trabaja con todo lo disponible. Por esta razón, se intentará construir una base de datos fiable, no redundante y lo más extensa posible.

Hipótesis

Los virus están expuestos a sesgos mutacionales propios del hospedero. Aquellos virus más expuestos a estos sesgos, presentarán efectos más evidentes.

Se espera que utilizando estadística multivariada sea posible agrupar virus por sus firmas genómicas y por sus hospederos.

Objetivo general

Explorar la diversidad viral e intentar describir grandes patrones evolutivos.

Objetivos específicos

Construir una base de datos no redundante de virus.

Realizar análisis composicionales de genomas virales completos.

Determinar si existen virus que presentan usos de dinucleótidos particulares.

MATERIALES Y MÉTODOS

Construcción de la base de datos

Las secuencias de los virus se obtuvieron de la base de datos *Viral Genomes*, base secundaria del NCBI (del inglés, *National Center for Biotechnology Information*). Las descargas se hicieron desde su sitio ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/ [11]. *Viral Genomes* contaba con 4375 entradas al momento de entregar este trabajo. Debido a que se optó por dejar de lado a los virus satélites, se trabajó con 4195 virus.

Utilizando la línea de comandos y *scripts* en lenguaje bash, se obtuvo información de los archivos disponibles en *Viral Genomes*. El tipo de virus, estructura, taxonomía, hospedero, etc., se obtuvieron de los archivos con formato GenBank (.gbk). Otras características como tamaño del genoma, frecuencia de bases, contenido de G y C y frecuencia de dinucleótidos, fueron obtenidos de los archivos que presentan el genoma en formato FASTA (.fna). Algunos virus no tenían asociado un archivo .fna y se construyeron estos archivos con la secuencia del genoma incluida en el archivo .gbk (ver **ANEXO Cuadro A1**).

Debido a que los archivos .gbk incluían solamente un ítem de taxonomía, aquellos virus asignados a un orden, presentan ese dato. Aquellos que pertenecen a una familia no asignada, presentan la familia a la cual pertenecen. En el caso de no estar asignado a una familia, se incluye el género en el que se los clasifica. Un número grande de virus aparece como sin clasificar. Esto se debe a que algunos virus no han sido muy estudiados o a que su taxonomía está en discusión, quedando pendiente la resolución de su clasificación por el ICTV. Para poder tener un criterio de clasificación común a todos los virus, se completó la familia a la que pertenecen en base a la clasificación del ICTV. Para disminuir el número de virus sin asignar, se buscó información en la publicación de referencia de los genomas con poca información.

Análisis composicional

Para cada virus disponible se determinó su composición de bases y sus frecuencias de dinucleótidos, a partir de los archivos .fna.

Se calcularon las frecuencias relativas de cada dinucleótido, y para cada virus, utilizando la fórmula propuesta por Samuel Karlin:

$$\rho^*_{XY} = \frac{f_{XY}}{f_X . f_Y}$$

Para evidenciar sesgos en las frecuencias de dinucleótidos se observarán aquellos valores menores o iguales a 0,78 y aquellos mayores o iguales que 1,23 [5].

Análisis multivariado

Se generó una matriz de 16 columnas (4^2 dinucleótidos posibles) y 4195 filas (tantas filas como virus). Con esta matriz se hizo un análisis de componentes principales (PCA, del inglés *principal component analysis*) en R con el paquete FactoMineR [**13**,**14**]. Se incluyeron también variables suplementarias, tanto cuantitativas como cualitativas. Las cuantitativas fueron el tamaño y frecuencias de bases en el genoma. Las cualitativas fueron diferentes características de los virus, como el tipo de genoma, su estructura, el grupo de Baltimore al que pertenecen, orden y/o familia en que el ICTV lo incluye y el grupo de hospedero al que parasitan. Se estudiarán los principales ejes y se representarán gráficamente los virus en estos ejes, coloreándolos según las variables cualitativas.

RESULTADOS

Diversidad de los genomas virales disponibles en la base de datos *Viral Genomes*

Se analizó un total de 4195 virus. Un 64% de ellos poseía un genoma de ADN; el 36% restante poseía genoma de ARN. En cuanto a su estructura, un 52% del total fueron de hebra doble y el otro 48% fueron de hebra simple. Al observar a qué grupo de Baltimore pertenecía cada uno de los virus, algunos grupos se encontraron muy representados y otros muy poco, como ambos grupos retrotranscritos (ver **Cuadro 2**).

Cuadro 2. Proporción de virus pertenecientes a cada uno de los grupos de Baltimore (ordenados del grupo *I* al *VII*).

Grupo	%
dsDNA	45
ssDNA	17
dsRNA	5
ssRNA(+)	23
ssRNA(-)	6
ssRNA(RT)	2
dsDNA(RT)	2

ICTV tiene aceptadados siete órdenes; el número de virus asignados a estos, según su ultimo listado, es de 1052. En la base de datos *Viral Genomes* son 1900. Quedan sin asignar, es decir, sin orden asignado, 2134 y 2295 virus, respectivamente (ver **Cuadro 3**).

Cuadro 3. Comparación del número de virus por orden disponibles en Viral Genomes y los clasificados en ICTV.

Orden	*Viral Genomes*	ICTV
Caudovirales	1289	456
Picornavirales	165	152
Mononegavirales	155	123
Tymovirales	146	176
Herpesvirales	66	102
Nidovirales	66	31
Ligamenvirales	13	12
sin asignar	2295	2134
	4195	**3186**

Se reunió información de cada familia representada (ver **ANEXO Cuadro A2**). Casi todas se encuentran aceptadas por ICTV, siendo la única excepción la familia *Mycodnaviridae,* pendiente de aprobación [**14**]. Entre los 4195 virus, hay siete familias designadas por ICTV que no están representadas (ver **Cuadro 4**).

Cuadro 4*. Comparación del número de virus por familia disponibles en *Viral Genomes* y los clasificados en ICTV. *Mycodnaviridae*, en verde, no ha sido aceptada aún por ICTV [**14**].

* ver página siguiente

Familia	*Viral Genomes*	ICTV
Siphoviridae	678	313
Geminiviridae	368	325
Myoviridae	300	93
Podoviridae	219	50
Papillomaviridae	128	95
Potyviridae	123	190
Parvoviridae	84	56
Picornaviridae	83	50
Rhabdoviridae	80	71
Flaviviridae	78	60
Betaflexiviridae	70	87
Baculoviridae	68	49
Polyomaviridae	65	13
Retroviridae	63	53
Circoviridae	60	12
Paramyxoviridae	60	36
Caulimoviridae	58	53
Reoviridae	58	87
Herpesviridae	57	88
Tombusviridae	57	71
Adenoviridae	51	44
Coronaviridae	50	25
Virgaviridae	48	54
Anelloviridae	46	66
Alphaflexiviridae	45	51
Totiviridae	43	28
Bunyaviridae	42	100
Secoviridae	41	73
Inoviridae	37	43
Partitiviridae	37	56
Poxviridae	37	69
Astroviridae	36	22
Closteroviridae	35	39
Arenaviridae	34	30
Bromoviridae	34	33
Tymoviridae	30	37
Luteoviridae	29	33
Caliciviridae	24	7
Togaviridae	24	32
Mycodnaviridae [14]	23	0
Iflaviridae	19	9
Microviridae	19	12
Dicistroviridae	18	15
Endornaviridae	17	8
Iridoviridae	16	11
Phycodnaviridae	16	33
Nodaviridae	15	9
Narnaviridae	13	7
Hepadnaviridae	12	7
Leviviridae	11	4
Orthomyxoviridae	11	8
Fuselloviridae	10	9
Hypoviridae	10	4

Familia	*Viral Genomes*	ICTV
Arteriviridae	9	4
Birnaviridae	8	6
Filoviridae	8	7
Lipothrixviridae	8	9
Nanoviridae	8	10
Alloherpesviridae	7	12
Mesoniviridae	6	1
Chrysoviridae	5	9
Cystoviridae	5	1
Mimiviridae	5	2
Polydnaviridae	5	53
Rudiviridae	5	3
Tectiviridae	5	5
Ascoviridae	4	4
Hepeviridae	4	5
Nudiviridae	4	3
Nyamiviridae	4	4
Ophioviridae	4	6
Sphaerolipoviridae	4	6
Alphatetraviridae	3	10
Amalgaviridae	3	4
Benyviridae	3	4
Bicaudaviridae	3	1
Bornaviridae	3	5
Marseilleviridae	3	4
Globuloviridae	2	2
Hytrosaviridae	2	2
Malacoherpesviridae	2	2
Turriviridae	2	2
Alvernaviridae	1	1
Ampullaviridae	1	1
Asfarviridae	1	1
Barnaviridae	1	1
Bidnaviridae	1	1
Carmotetraviridae	1	1
Corticoviridae	1	1
Gammaflexiviridae	1	1
Marnaviridae	1	1
Megabirnaviridae	1	1
Nimaviridae	1	1
Permutotetraviridae	1	2
Picobirnaviridae	1	2
Plasmaviridae	1	1
Quadriviridae	1	1
Roniviridae	1	1
Avsunviroidae	0	4
Clavaviridae	0	1
Guttaviridae	0	2
Metaviridae	0	39
Pospiviroidae	0	28
Pseudoviridae	0	34
Spiraviridae	0	1
sin asignar	355	43
	4195	**3186**

Aquellos virus sin familia asignada fueron categorizados dentro de alguno de los grupos o según la composición de su genoma (ver **Cuadro 5**). Un 94% pudo ser incluido en alguna categoría taxonómica (i.e., *Caudovirales*, *Picornavirales*) o según la clasificación de Baltimore. Menos de 20 virus debieron categorizarse por su genoma (ADN o ARN). Solamente un virus quedó sin agrupar en ninguna categoría: el "virus" *Planaria asexual strain-specific virus-like element type 1*.

Cuadro 5. Virus sin familia asignada; se indica la categoría en la que se los incluyó (orden, grupo de Baltimore, estructura o tipo de genoma).

Virus sin familia asignada	Nº
Caudovirales sin clasificar	92
ssDNA virus	70
ssRNA(+) virus	60
dsDNA phage	50
dsDNA virus	33
dsRNA virus	18
DNA phage	11
ssRNA(-) virus	7
RNA virus	5
DNA archaeal virus	3
Picornavirales sin clasificar	3
dsDNA archaeal virus	2
sin asignar	1
	355

Se analizó la proporción de grupos de hospederos existente entre estos virus. Bacterias y animales son los que más virus tuvieron en esta muestra; cada uno un 35%. Otro porcentaje importante, un 25%, fueron virus de plantas. Los restantes grupos de hospederos tuvieron muy pocos de sus virus totalmente secuenciados; 2% de hongos y de arqueas y 1% de algas y de protozoarios (ver **Cuadro 6**).

Cuadro 6. Grandes grupos de hospederos y el número de virus que parasita cada uno de estos grupos.

Hospedero	Nº de virus
Bacterias	1426
Animales	1422
Plantas	1022
Hongos	88
Arqueas	69
Algas	41
Protozoarios	26

Un 38% de los genomas virales no tenía indicada la especie a la que parasitan. En el 62% restante aparecía el nombre vernáculo en algunos, género y especie en otros o solamente el género; muy pocos virus poseían información más detallada (e.g., cepa, subespecie). Se contaron los géneros que más aparecían representados como hospederos (ver **Cuadro 7**). Luego del hombre, aparecen tres géneros de bacterias. A continuación, apareció un género vegetal (*Solanum*). De los primeros doce, nueve eran bacterianos. *Sulfolobus* y *Acidianus*, dos géneros de arqueas, tenían descritos 15 y 11 virus, respectivamente.

Cuadro 7. Género de hospederos más representados (10 o más virus) y el número de virus parásito de cada género.

Género de hospedero	Nº de virus
Homo	114
Escherichia	110
Pseudomonas	87
Staphylococcus	75
Solanum	74
Bacillus	56
Sus	38
Vibrio	38
Streptococcus	36
Burkholderia	30
Mycobacterium	30
Salmonella	29
Bos	23
Gallus	22
Lactococcus	22
Propionibacterium	21
Clostridium	20
Cellulophaga	18
Culex	18
Equus	18
Listeria	17
Synechococcus	16
Vitis	16
Aeromonas	15
Enterococcus	15
Sulfolobus	15
Allium	14
Canis	14
Acinetobacter	13
Cronobacter	12
Mus	12
Ralstonia	12
Yersinia	12
Acidianus	11
Felis	11
Meleagris	11
Xanthomonas	10

Se determinó el código genético utilizado en la transcripción para cada familia: todas utilizan el código universal. Algunos casos particulares son *Inoviridae*, *Microviridae*, *Narnaviridae* y *Podoviridae* (ver **ANEXO Cuadro A2**).

Frecuencias relativas de dinucleótidos

Se obtuvieron los valores de frecuencia relativa de cada uno de los dieciséis dinucleótidos, para cada uno de los virus, y se calcularon medidas de tendencia central para el total de los virus y para cada una de las categorías tenidas en cuenta previamente

(i.e., tipo de genoma, hebra doble o hebra simple, grupo de Baltimore, orden, familia, hospedero). Para el total de los virus, únicamente se observó sub-representación del dinucleótido TpA (ver **Cuadro 8**).

Cuadro 8. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) para el conjunto de todos de los virus; en rojo se señala el valor menor a 0,78.

Todos (n)	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
(4195)	1,07	0,98	0,98	0,97	1,11	1,02	0,82	1,02	1,07	1,01	1,01	0,93	0,76	1,02	1,15	1,07

Para los virus con genoma de ADN se obtuvo sub-representación de TpA, mientras que los ARN, además del TpA, tuvieron sub-representado el dinucleótido CpG y presentaron sobrerrepresentación de TpG (ver **Cuadro 9**).

Cuadro 9. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) para virus con genoma de ADN y de ARN; en rojo se señalan los valores menores o iguales a 0,78 y en verde los mayores o iguales a 1,23.

Genoma (n)	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
ADN (2689)	1,08	0,97	0,96	0,98	1,07	1,01	0,90	1,00	1,06	1,03	1,00	0,94	0,78	1,02	1,11	1,07
ARN (1506)	1,06	0,98	1,01	0,95	1,17	1,03	0,67	1,06	1,10	0,97	1,03	0,90	0,72	1,02	1,23	1,08

Los virus de hebra doble tuvieron una sub-representación de TpA, mientras que los de hebra simple la tuvieron tanto de TpA como CpG (ver **Cuadro 10**).

Cuadro 10. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) para virus de hebra doble (ds) y de hebra simple (ss); en rojo se señalan los valores menores a 0,78.

Hebra (n)	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
ds (2168)	1,07	1,00	0,96	0,98	1,09	0,96	0,91	1,00	1,06	1,06	0,97	0,94	0,76	1,02	1,11	1,08
ss (2010)	1,08	0,96	0,99	0,97	1,13	1,09	0,71	1,04	1,09	0,96	1,04	0,91	0,75	1,02	1,20	1,07

Con la excepción del grupo *II*, todos los demás grupos de virus presentaron sub-representación de al menos uno de los dinucleótidos. De estos seis, solamente el *VI* no mostró sub-representación de TpA y los grupos *I* y *III* de CpG (ver **Cuadro 11**).

Cuadro 11. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) por grupo de Baltimore ordenados de *I* a *VII*; en rojo se señalan los valores menores o iguales a 0,78 y en verde los mayores o iguales a 1,23.

Grupo (n)	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
dsDNA (1891)	1,07	1,00	0,96	0,98	1,09	0,95	0,93	1,00	1,05	1,07	0,97	0,95	0,76	1,02	1,11	1,08
ssDNA (716)	1,10	0,93	0,94	1,01	1,02	1,15	0,84	1,00	1,07	0,95	1,05	0,93	0,82	1,00	1,14	1,05
dsRNA (207)	1,04	1,04	0,92	1,00	1,11	0,94	0,91	1,01	1,11	1,00	0,95	0,93	0,78	1,01	1,19	1,04
ssRNA(+) (978)	1,08	1,00	1,00	0,93	1,18	1,03	0,70	1,05	1,08	0,99	1,01	0,92	0,70	0,99	1,24	1,09
ssRNA(-) (252)	1,00	0,90	1,10	1,01	1,24	1,08	0,41	1,12	1,15	0,88	1,10	0,83	0,70	1,16	1,24	1,02
ssRNA(RT) (63)	1,08	0,89	1,11	0,92	1,05	1,23	0,51	1,17	1,03	0,93	1,25	0,82	0,83	0,96	1,14	1,09
dsDNA(RT) (70)	1,04	0,86	1,14	0,95	1,12	1,14	0,51	1,16	1,13	0,94	1,09	0,78	0,73	1,15	1,12	1,12

Algunos de los grupos mostraron sobrerrepresentación de dinucleótidos: *IV* y *V* de TpG, *V* de CpA y *VI* tanto de GpG como CpC (ver **Cuadro 11**).

En los órdenes virales se observaron varios dinucleótidos sobrerrepresentados, mientras que solamente TpA y CpG estuvieron sub-representados en alguno de ellos; TpA en *Caudovirales*, *Mononegavirales*, *Picornavirales* y *Tymovirales*; CpG en *Mononegavirales*, *Nidovirales* y *Picornavirales*. Los órdenes con sobrerrepresentación de dinucleótidos fueron *Picornavirales* (TpG), *Tymovirales* (ApA, GpA y TpC) y *Nidovirales* (CpA y TpG) (ver **Cuadro 12**). El mismo abordaje se aplicó en las familias; TpA fue el más veces sub-representado (en 49 de las 98), seguido de CpG (en 38); sobrerrepresentación de TpG (en 23) y de CpA (en 15) (ver **ANEXO Cuadro A3**).

Cuadro 12. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) por orden viral; en rojo se señalan los valores menores a 0,78 y en verde los mayores a 1,23.

Orden (n)	**ρ*AA**	**ρ*AC**	**ρ*AG**	**ρ*AT**	**ρ*CA**	**ρ*CC**	**ρ*CG**	**ρ*CT**	**ρ*GA**	**ρ*GC**	**ρ*GG**	**ρ*GT**	**ρ*TA**	**ρ*TC**	**ρ*TG**	**ρ*TT**
Caudovirales (1289)	1,06	1,00	0,95	0,99	1,08	0,91	0,97	1,00	1,07	1,08	0,94	0,96	0,74	1,05	1,10	1,06
Picornavirales (165)	1,11	1,00	0,97	0,92	1,20	1,15	0,49	1,09	1,10	0,93	1,11	0,89	0,66	0,93	1,36	1,07
Mononegavirales (155)	0,97	0,90	1,08	1,04	1,21	1,10	0,47	1,11	1,18	0,84	1,12	0,82	0,72	1,18	1,21	1,01
Tymovirales (146)	1,26	0,86	1,16	0,87	1,12	0,94	0,79	1,12	1,26	0,92	1,12	0,84	0,41	1,26	1,10	1,03
Herpesvirales (66)	1,11	0,97	0,97	0,96	1,04	1,03	0,94	0,97	1,02	1,01	1,02	0,96	0,82	1,01	1,04	1,12
Nidovirales (66)	1,04	1,12	0,97	0,92	1,25	1,02	0,58	1,06	0,90	1,16	0,94	1,03	0,86	0,80	1,33	1,01
Ligamenvirales (13)	1,05	0,97	1,01	0,95	1,08	0,99	0,90	0,97	1,05	0,98	1,01	0,97	0,88	1,06	1,02	1,08

Por grupo de hospedero no se observó ninguna sobrerrepresentación. TpA y CpG se encontraron sub-representados en varios de ellos: TpA para todos excepto en arqueas y CpG para animales y plantas (ver **Cuadro 13**).

Cuadro 13. Medias de las frecuencias relativas de dinucleótidos (**ρ***NN) por grupo de hospedero, incluidos los que no tienen resuelto su hospedero entre un invertebrado o una planta (invert./plantas) y los genomas producto de metagenómica de hospedero desconocido (*"environment"*); en rojo se señalan los valores menores o iguales a 0,78.

Hospedero (n)	**ρ*AA**	**ρ*AC**	**ρ*AG**	**ρ*AT**	**ρ*CA**	**ρ*CC**	**ρ*CG**	**ρ*CT**	**ρ*GA**	**ρ*GC**	**ρ*GG**	**ρ*GT**	**ρ*TA**	**ρ*TC**	**ρ*TG**	**ρ*TT**
bacterias (1426)	1,07	1,00	0,95	0,99	1,08	0,91	0,97	1,00	1,06	1,09	0,95	0,95	0,74	1,04	1,10	1,06
animales (1422)	1,06	0,99	1,00	0,95	1,15	1,08	0,68	1,04	1,04	0,99	1,07	0,91	0,77	0,97	1,20	1,09
plantas (1022)	1,08	0,93	0,99	0,98	1,10	1,10	0,76	1,02	1,11	0,96	1,00	0,91	0,75	1,04	1,18	1,08
hongos (88)	1,06	1,04	0,94	0,97	1,08	0,97	0,89	1,05	1,09	0,97	1,01	0,94	0,78	1,03	1,14	1,04
arqueas (69)	1,03	1,05	0,97	0,93	0,95	0,95	1,03	1,02	1,16	0,93	0,99	1,00	0,82	1,15	0,96	1,02
invert./plantas (58)	1,11	0,89	1,10	0,97	1,17	0,98	0,64	1,10	1,21	0,92	1,06	0,85	0,60	1,20	1,18	1,01
"*environment*" (43)	1,07	0,94	0,94	1,03	1,04	1,07	0,90	0,97	1,10	0,90	1,03	0,99	0,81	1,08	1,11	1,01
algas (41)	1,10	0,98	0,90	1,01	1,08	1,01	1,01	0,92	1,09	0,92	1,01	0,97	0,75	1,08	1,09	1,10
protozoarios (26)	1,10	1,00	0,91	0,96	1,11	1,03	0,90	0,96	1,07	0,97	1,07	0,92	0,77	1,03	1,12	1,12

Se realizó un PCA con los valores de frecuencia relativa de dinucleótidos de todos los genomas analizados (matriz de 4195 por 16) (ver **Figura 2**).

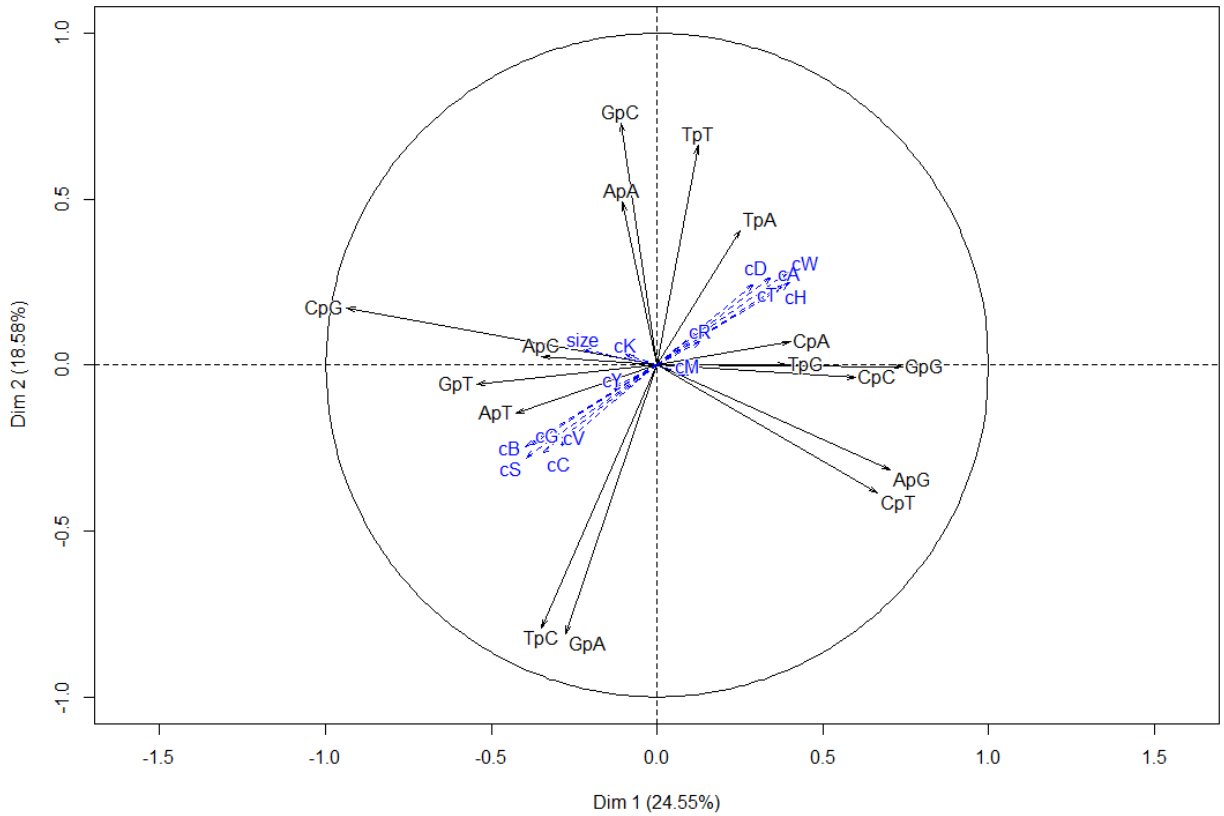

Figura 2. Análisis de componentes principales para las frecuencias relativas de dinucleótidos (**NpN**) (flechas negras); en azul (flechas punteadas) vectores de variables cuantitativas suplementarias (size: tamaño del genoma; cA: contenido de adenina; cB: cualquiera excepto A; cC: contenido de citosina; cD: cualquiera excepto C; cG: contenido de guanina; cH: cualquiera excepto G; cK: contenido de G y T (timina); cM: contenido de A y C; cR: contenido de A y G (purinas); cS: contenido de G y C; cT: contenido de T); cV: cualquiera excepto T; cW: contenido de A y T; cY: contenido de C y T (pirimidinas)).

El PCA dio como resultado muchos ejes con una proporción relativamente grande de varianza explicada. Cuatro ejes presentaron una varianza explicada mayor al 10%. Para explicar más de un 90% y de un 95% se debe recurrir a 8 y 9 ejes, respectivamente (ver **Figura 3**).

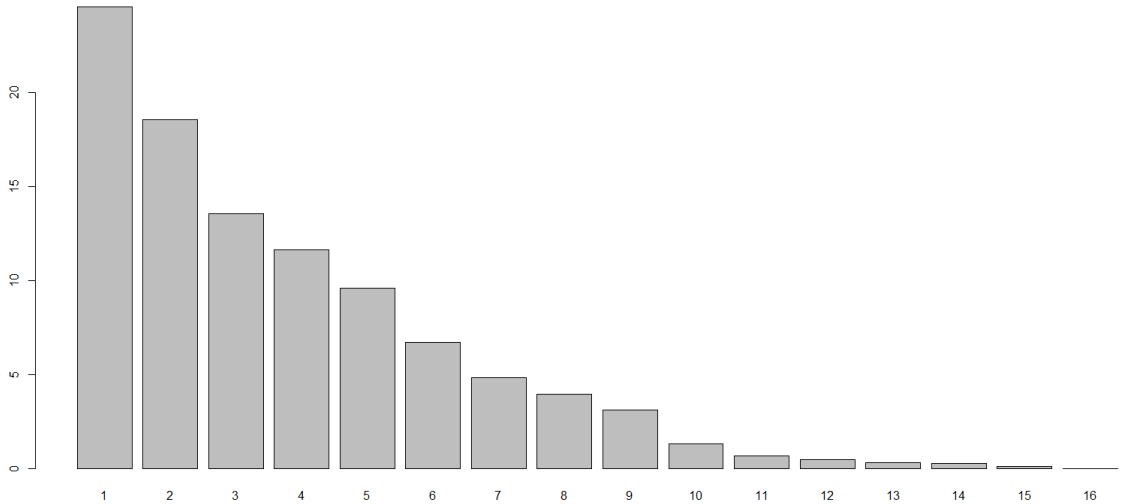

Figura 3. Varianza explicada (en %) por cada uno de los dieciséis ejes para las frecuencias relativas de dinucleótidos.

Se estudió la correlación entre los primeros cuatro ejes principales del PCA y cada una de las variables cuantitativas (ver **Cuadro 14**). Para las frecuencias relativas de dinucleótidos, variables utilizadas en el análisis multivariado, el primer eje presentó una fuerte correlación negativa con los valores de CpG (ver **Cuadro 14. A**). Para las variables suplementarias, este eje presentó correlación media con prácticamente todas las bases, excepto T, presentando correlación baja para el contenido de T y de G y C (y de A y T) y para el tamaño del genoma (ver **Cuadro 14. B**). El cuarto eje se correlacionó negativamente con TpA y positivamente con CpA y TpG (ver **Cuadro 14. A**).

Cuadro 14. A: Correlación entre cada uno de los primeros cuatro ejes principales (**Dim.1**, **Dim2**, **Dim.3** y **Dim.4**) y las frecuencias relativas de dinucleótidos (**NpN**): en verde, correlaciones positivas; rojo, correlaciones negativas. **B:** Correlación entre los primeros cuatro ejes y las variables suplementarias cuantitativas; en verde y rojo, idem **A**. (size: tamaño del genoma; cA: contenido de adenina; cB: cualquiera excepto A; cC: contenido de citosina; cD: cualquiera excepto C; cG: contenido de guanina; cH: cualquiera excepto G; cK: contenido de G y T (timina); cM: contenido de A y C; cR: contenido de A y G (purinas); cS: contenido de G y C; cT: contenido de T); cV: cualquiera excepto T; cW: contenido de A y T; cY: contenido de C y T (pirimidinas)).

A

	Dim.1	Dim.2	Dim.3	Dim.4
ApA	-0,10	0,49	-0,59	0,12
ApC	-0,35	0,03	0,75	-0,17
ApG	0,70	-0,32	0,03	-0,13
ApT	-0,43	-0,15	-0,16	0,26
CpA	0,40	0,07	0,52	0,64
CpC	0,60	-0,04	-0,27	-0,31
CpG	-0,94	0,17	-0,15	-0,13
CpT	0,67	-0,38	0,06	-0,08
GpA	-0,28	-0,81	-0,28	0,14
GpC	-0,11	0,73	0,01	0,30
GpG	0,74	0,00	-0,21	-0,19
GpT	-0,54	-0,06	0,54	-0,37
TpA	0,25	0,41	0,17	-0,73
TpC	-0,35	-0,79	-0,30	0,12
TpG	0,39	0,00	0,43	0,58
TpT	0,12	0,66	-0,38	0,20

	Dim.1	Dim.2	Dim.3	Dim.4
size	-0,22	0,05	0,01	-0,06
cA	0,40	0,25	0,04	-0,09
cB	-0,40	-0,25	-0,04	0,09
cC	-0,34	-0,26	-0,03	0,12
cD	0,34	0,26	0,03	-0,12
cG	-0,38	-0,24	0,04	0,16
cH	0,38	0,24	-0,04	-0,16
cK	-0,09	0,03	-0,01	-0,04
cM	0,09	-0,03	0,01	0,04
cR	0,13	0,07	0,12	0,06
cS	-0,40	-0,28	0,00	0,15
cT	0,29	0,25	-0,05	-0,17
cV	-0,29	-0,25	0,05	0,17
cW	0,40	0,28	0,00	-0,15
cY	-0,13	-0,07	-0,12	-0,06

Se graficaron los primeros dos ejes principales (varianza acumulada de 43%) para casi todas las categorías. Los virus con genoma de ADN fueron más abundantes que los de ARN en tres de los cuatro cuadrantes (ver **Figura 4**). Esto no ocurrió entre los virus de hebra doble y hebra simple (ver **Figura 5**). Por grupo de Baltimore, presentaron una leve tendencia a agruparse (ver **Figura 6**). Lo mismo ocurrió con los virus pertenecientes a los distintos órdenes, aunque de manera más difusa (ver **Figura 7**). Cuando se los observó por hospedero, los tres grupos con mayor representación (bacterias, animales y plantas), se ubicaron relativamente agrupados (ver **Figura 8**).

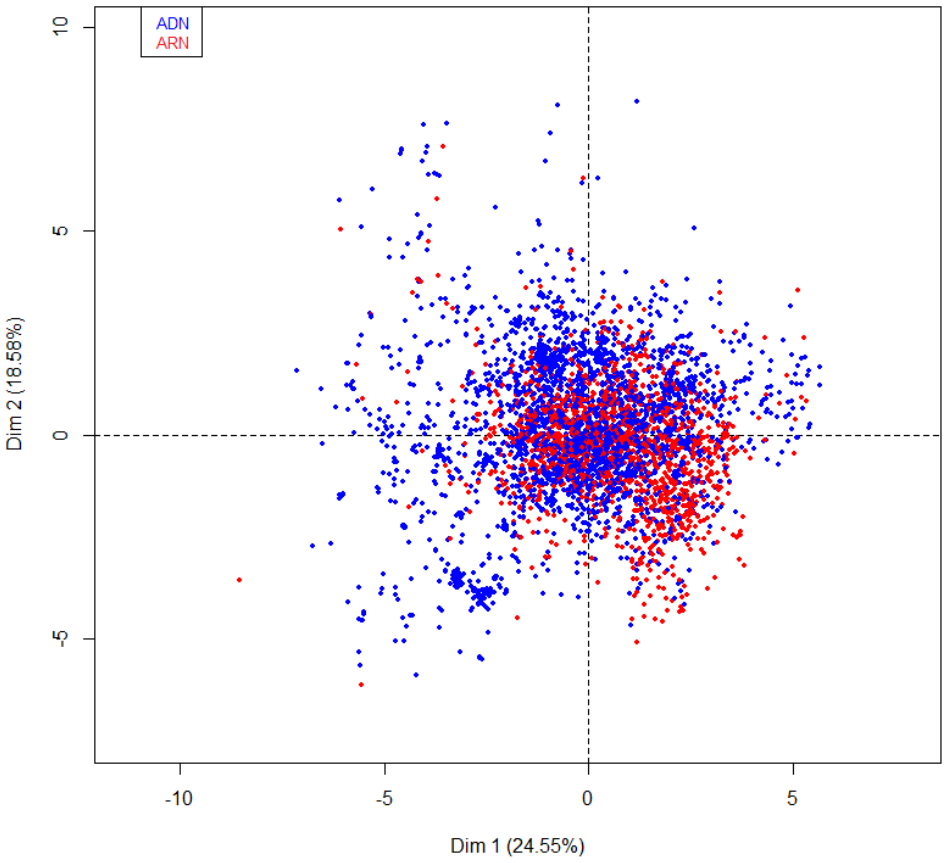

Figura 4. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por tipo de genoma (ADN o ARN).

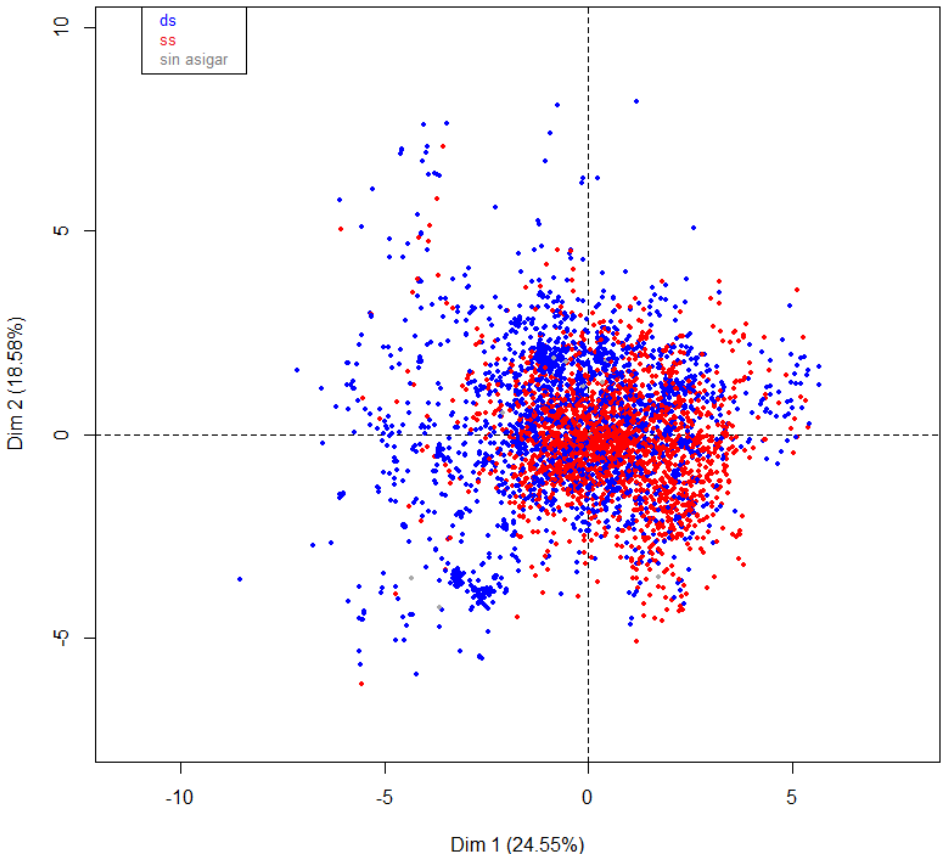

Figura 5. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por si su estructura genómica es de hebra doble (ds) o hebra simple (ss).

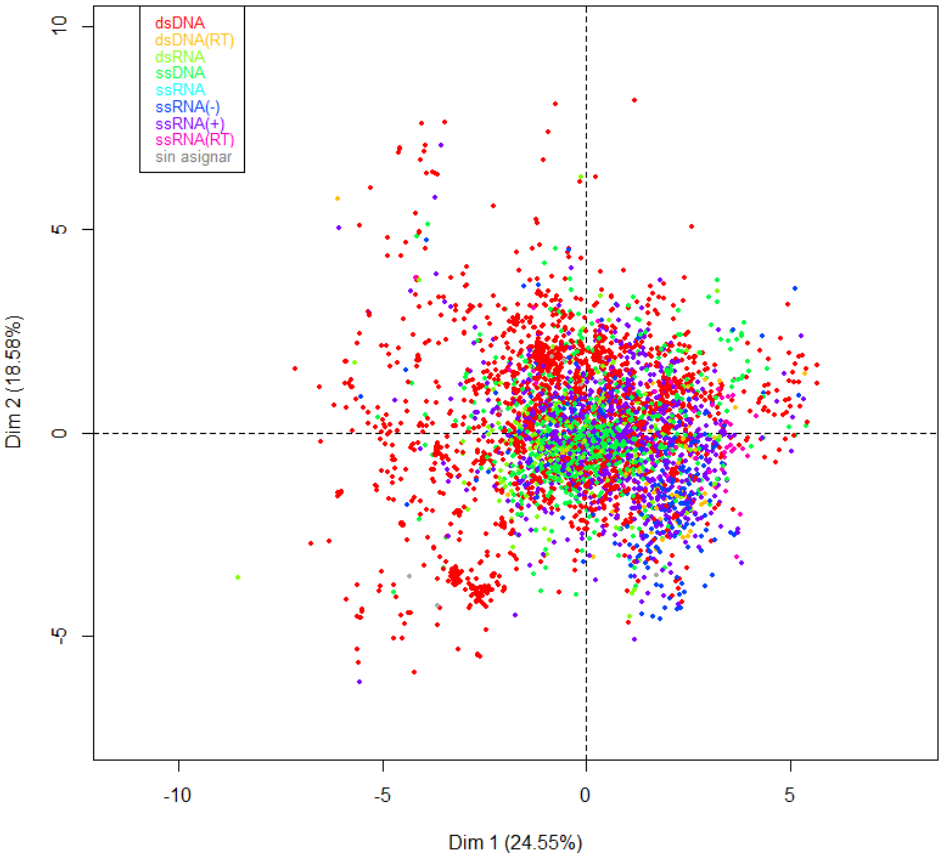

Figura 6. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por grupo de Baltimore; en turquesa, virus de ARN de hebra simple y de polaridad desconocida (ssRNA).

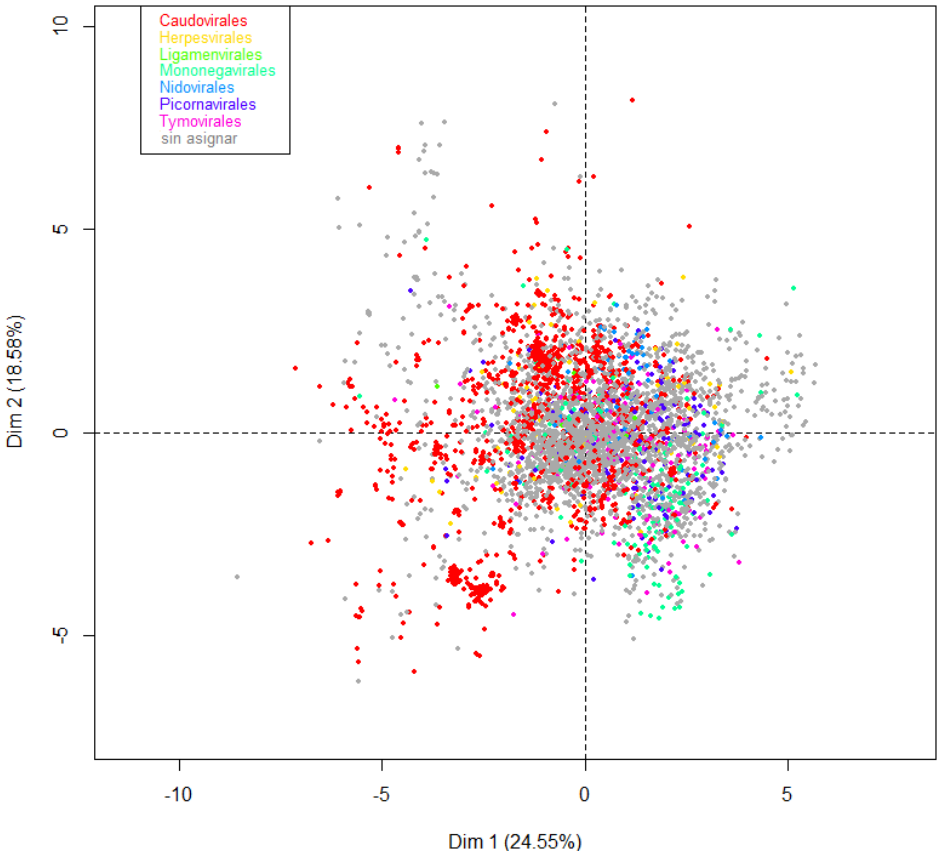

Figura 7. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por orden taxonómico del ICTV.

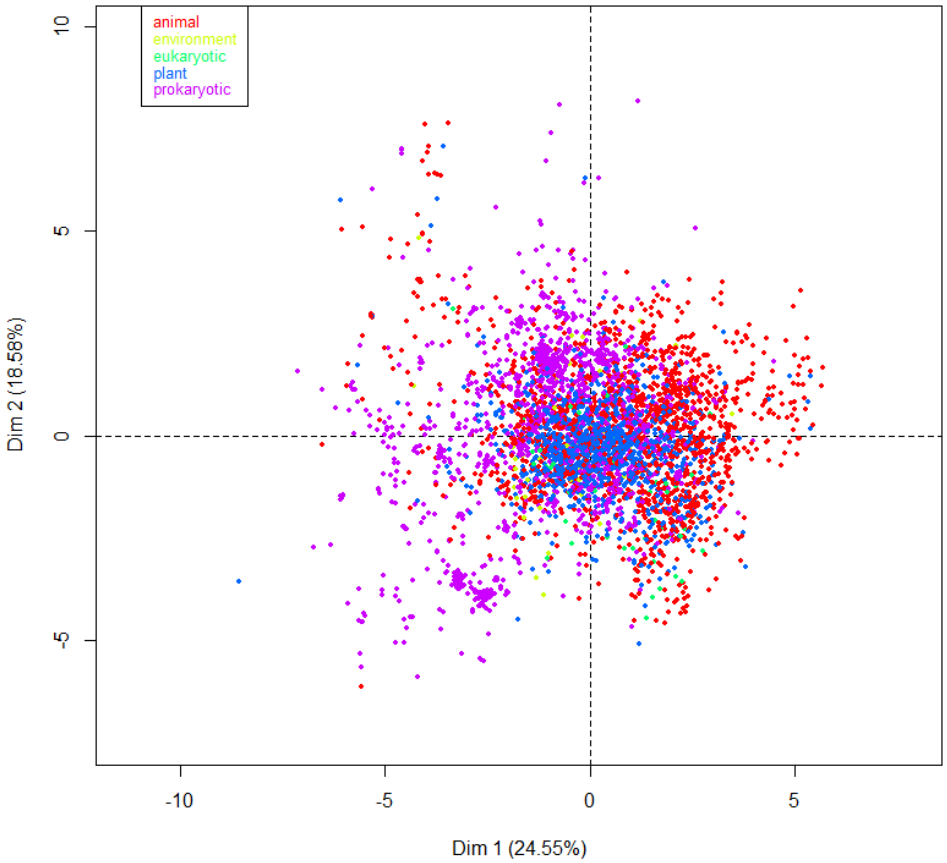

Figura 8. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por grupo de hospedero; environment refiere a genomas de metagenómica de hospedero desconocido; eukaryotic refiere a virus encontrados en insectos fitófagos, sin resolver si son virus de estos insectos o de las plantas.

DISCUSIÓN

Diversidad viral

En *Viral Genomes* existe una diversidad viral muy grande. Esto es lo esperable para las entidades más abundantes de la naturaleza. Que exista un número considerablemente mayor de genomas de virus de ADN (casi 2:1 con los ARN) se debe a que la gran mayoría de los bacteriófagos son dsDNA (es decir, grupo *I* de Baltimore). Entre hebra doble y hebra simple no se observan grandes diferencias en los porcentajes; aunque el grupo *I* es el predominante, entre grupo *IV* y grupo *II* reúnen un 40% del total.

Las diferencias encontradas en el número de especies dentro de los órdenes entre *Viral Genomes* y el ICTV pueden deberse a dos razones, no excluyentes pero opuestas. El ICTV se organiza formando comités y proponiendo cambios en la taxonomía; esto ocasiona que la designación de nuevas especies virales se demore, o al menos avance más lentamente que la secuenciación de nuevos genomas. Esto es particularmente evidente con metagenómica de ambientes donde los bacteriófagos se encuentran en un orden de magnitud mayor que las bacterias.

La otra razón podría explicar que se disponga de casi tres veces más de genomas de los *Caudovirales* en NCBI que en ICTV es a la existencia de genomas redundantes. Resultados (no presentados) de *clustering* por similitud de secuencia dieron que existe una proporción muy baja de secuencias redundantes (menos de 2% de secuencias "idénticas" en *Caudovirales*, pero con algunos casos con un 99 o 100% de identidad).

16 de las familias tienen un único representante y 38 de las familias no tienen más de cinco. Además hay 355 virus sin asignar. Algunos virus que no presentan ni orden ni familia, puede que formen parte de algún género que no esté aún bien clasificado. El género *Pandoravirus*, compuesto solamente por *P. salinus* y *P. dulcis*, casualmente los virus de genoma más grande conocido, está en esta situación.

Planaria asexual strain-specific virus-like element type 1 fue el único genoma que quedó sin agrupar en ninguna categoría; *"virus-like element"* parece indicar que no es un virus propiamente dicho.

Entre las siete familias propuestas por ICTV que no quedaron representadas en esta base de datos, tres tienen un número considerable de especies (ver **Cuadro 4)**: *Metaviridae* (39), *Pseudoviridae* (34) y *Pospiviroidae* (28). Estas familias pertenecen al grupo *VI*, el de los ssRNA(RT). De incluirse en la base de datos (suponiendo que sus genomas están secuenciados), en conjunto supera el porcentaje de virus asignados al *VI*, por lo tanto más que se duplicaría su número. Incluyendo a estos virus, se tendría una mayor paridad entre el número de virus con genoma de doble hebra y de simple hebra. Que no estén en la base de datos construida puede deberse a que se los filtró al momento de deshacerse de los virus satélites. También puede haber pasado que éstos estén categorizados en la base de datos generado como virus de grupo sin asignar o como ssRNA de polaridad desconocida (ver **Figura 6**). Queda como perspectiva revisar que sucedió con estas tres familias de virus y con las otras que no se incluyeron.

Los hospederos pueden dividirse en tres grandes grupos y otros grupos menores. No se incluyeron en el cálculo de los porcentajes de grupos de hospederos a algunas familias de virus con un amplio espectro de hospederos: miembros de las familias *Bunyaviridae*, *Reoviridae*, *Rhabdoviridae* y *Tymoviridae*, 58 virus en total. En cada una de estas cuatro familias, algunos de sus miembros parasitan plantas y otros parasitan invertebrados. Tampoco se incluyeron en los cálculos a los virus obtenidos por metagenómica ambiental, más exactamente a aquellos secuenciados con dicha técnica y que no lograron asignarle hospedero; *environment* es el hospedero que figura en el NCBI en estos casos.

Se presentó el código genético utilizado en la traducción por cada familia viral (ver **ANEXO Cuadro A2**). Esta característica es consecuencia directa de su hospedero. La gran mayoría utiliza el código genético universal (1 y 11). *Plasmaviridae* figura con el código genético 1, aunque es una familia de bacteriófagos; seguramente sea un error de quien subió el genoma, pero dado que en los codones no hay diferencias (exceptuando en los codones de inicio alternativo), no es un error de importancia. Cuatro de las familias tienen miembros que usan un código genético alternativo (4). *Inoviridae*, *Microviridae*, y *Podoviridae* son bacteriófagos. *Narnaviridae* son virus de hongos. Ambos grupos de hospederos se encuentran descritos entre los que utilizan este código genético [**15**].

Sesgos en dinucleótidos

El dinucleótido TpA presentó importantes sesgos, en muchas de las categorías. Algunas excepciones fueron los grupos *II* y *VI*, los órdenes *Herpesvirales*, *Ligamenvirales* y *Nidovirales*, los virus de arqueas y los virus asignados a *environment*. Este trabajo intentó simplificar los niveles de sesgo. Karlin y Burge [6] describen niveles intermedios: entre 0,79 y 0,82 y entre 1,20 y 1,22 como marginalmente bajos y marginalmente altos, respectivamente; y entre 0,83 y 1,19 como producto del azar. Considerando estos niveles intermedios, solamente el grupo *VI* y los órdenes *Ligamenvirales* y *Nidovirales* no presentaron sesgo para TpA.

Ligamenvirales es un orden de virus de arqueas (grupo que presentó un sesgo marginalmente bajo); más precisamente de arqueas termófilas. La única familia con sobrerrepresentación de TpA fue la familia *Globuloviridae* (ver **ANEXO Cuadro A3**), que aunque no pertenece a *Ligamenvirales*, tiene como hospedero a arqueas termófilas. Virus de arqueas termófilas presentan contenidos de G y C marcadamente más bajos que los de bacterias termófilas [**16**]. Es posible que estos contenidos de T y A más altos sean un freno para el empobrecimiento en TpA en los *Ligamenvirales* pero en *Globuloviridae* potencie el enriquecimiento observado.

Nidovirales es un orden de virus de animales, que pertenecen al grupo *IV*. Algunos *Coronaviridae* superan los 30 kb, lo que los ubica entre los genomas de ARN más grandes. Se ha propuesto que estos virus comenzaron a expandir su tamaño después de incorporar un dominio con actividad exonucleasa 3'→5' en su ARN polimerasa, pudiendo operar como mecanismo de corrección de pruebas (*proofreading*) [**17**]. *Herpesvirales* (0,82), virus de vertebrados del grupo *I*, con tamaños entre 100 y 300 kb, tiene descrito un dominio con actividad *proofreading* en su ADN polimerasa [**18**]. Presentando esta actividad, muchos cambios espontáneos que ocurran serán corregidos.

Se observó sesgo en CpG en los virus de ARN, pero no en los virus de ADN, pero seguramente sea consecuencia de la gran cantidad de bacteriófagos representados. Los virus de ARN presentaron sesgo en TpG, pero no en CpA (1,17), aunque presentó el segundo valor más alto, luego de TpG.

Los virus de hebra simple presentaron sesgo de CpG, pero no los de hebra doble; de nuevo, esto quizá se deba al número de bacteriófagos representados en la muestra. Los tres grupos de virus de ARN de hebra simple, *IV*, *V* y *VI*, presentaron sesgos. También los del grupo *VII*; los dsDNA(RT) tienen una fase replicativa de ARN de hebra simple de polaridad positiva. Que los genomas de estos cuatro grupos estén en las células, en algún momento de su ciclo replicativo, como ARN de simple hebra, puede evidenciar que esta estructura es blanco de algún proceso molecular particular o también deberse a que estos genomas son los más propensos a sufrir mutaciones.

Los órdenes que presentaron sesgos bajos para CpG fueron *Mononegavirales*, *Nidovirales* y *Picornavirales*. *Nidovirales* infectan animales; *Mononegavirales* y *Picornavirales* animales y plantas. En estos tres órdenes existen familias que infectan invertebrados, donde no se esperaba encontrar sesgos en CpG, pero son minoritarias. *Tymovirale*s (0,79), virus de plantas, presentaron sesgo de CpG marginalmente bajo. *Nidovirales* también presentó valores altos de sesgo para TpG y para CpA. *Picornavirales* solamente presentó sesgo en TpG.

En los virus de animales y en los de plantas se observaron sesgos en CpG. También en los categorizados como de invertebrados o de plantas. Algunos de estos virus pueden encontrarse en insectos fitófagos, quedando por resolver si infectan a éstos o si son en realidad virus de las plantas de las que los insectos se alimentan. Además, algunos virus patógenos de plantas son arbovirus: son transmitidos por artrópodos. Entre estos virus con hospedero asignado de manera difusa, hay representantes de dos familias de arbovirus: *Bunyaviridae* y *Reoviridae*. Las restantes familias son *Tymoviridae*, del orden *Tymovirales*, típicamente de plantas, y *Rhabdoviridae*, del orden *Mononegavirales*, de plantas y animales.

Análisis de componentes principales

La varianza explicada por los primeros ejes principales fue relativamente baja. Al graficar estos ejes se observaron algunas tendencias, pero no claras agrupaciones. Las correlaciones entre variables y ejes principales, ya sean positivas o negativas, fueron disminuyendo su valor absoluto a medida que disminuía la varianza explicada por cada eje. El primer eje se asoció negativamente con CpG. Esto se observó tanto gráficamente (ver **Figura 2**) como estadísticamente (ver **Cuadro 14**).

Comparando los gráficos de genoma ADN y ARN con los de hebra doble y hebra simple, se observa como en el sector central predominan los ssDNA (ver **Figura 4** y **Figura 5**); se confirma observando la disposición por grupo de Baltimore (ver F**igura 6**).

Estos virus, los del grupo *II*, son predominantemente de plantas. Virus de procariotas y de animales ocupan más el plano formado por los ejes 1 y 2 (ver **Figura 8**).

El gráfico por orden resulta poco informativo; se observa la gran cantidad de virus sin orden asignado (ver **Figura 7**). Debido a la relativamente baja varianza explicada y a la gran cantidad de puntos, los gráficos que involucran más categorías se vuelven poco útiles, como ocurre con las familias (ver **ANEXO Figura A1**).

CONCLUSIONES

Las bases de datos presentan sesgos. Que existan cientos de virus de humanos y de *E. coli* lo confirma. Seguramente existan cientos (tal vez miles) de virus que infectan a estas especies, pero lo mismo puede formularse para cualquiera de las demás especies. Al contar con una cantidad de datos muy grande, y en continuo crecimiento, debería ir superándose en parte estos sesgos. La principal fuente de virus aún no caracterizados es la metagenómica.

Las debilidades a la hora de clasificarlos pueden presentar la oportunidad de generar un abordaje válido para lograr una mejor clasificación de los virus. Se pretendió agruparlos con estadística multivariada (en este caso, PCA). La posibilidad de formar grupos naturales se ve limitada por la gran variabilidad de estos, pero también por otros factores intrínsecos (e.g., genomas de gran tamaño, actividad *proofreading*) y/o extrínsecos. Aspectos que dependen del hospedero (e.g., sesgos mutacionales, respuesta del sistema inmune) también son determinantes.

Mientras el número de genomas de algas, arqueas, hongos y protozoarios no aumenten en algunos órdenes de magnitud, va a ser difícil extraer grandes patrones. Sobre todo si se analizan todos los virus en conjunto, donde existen tres grupos con más de mil virus en cada uno: bacterias, animales y plantas.

Se continuará trabajando sobre esta base de datos, pasando a analizar las secuencias codificantes. Resultados preliminares indican que a este nivel los virus presentan sesgos que permitirían resolver agrupamientos con mayor resolución.

Referencias

1. Edwards, R.A. & Rohwer, F. 2005. Viral metagenomics. Nature Reviews Microbiology 3(6):504-10

2. Baltimore, D. 1971. Expression of animal virus genomes. Bacteriological reviews 35(3):235-41

3. Hulo, C.; de Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. 2011. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Research 39(Database issue):D576-82. doi: 10.1093/nar/gkq901. Epub 2010 Oct 14

4. King, A.M.Q.; Adams, M.J.; Carstens, E.B.; Lefkowitz, E.J. (Eds.). 2012. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego, CA. ISBN 978-0-12-384684-6

5. Gentles, A.J. & Karlin, S. 2001. Genome-Scale Compositional Comparisons in Eukaryotes. Genome Research 11:540-46

6. Karlin, S. & Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics 11(7):283-90

7. Li, Q.; Li, N.; Hu, X.; Li, J.; Du, Z.; Chen, L.; Yin, G.; Duan, J.; Zhang, H.; Zhao, Y; Wang, J; Li, N. 2011. Genome-Wide Mapping of DNA Methylation in Chicken. PLoS One 6(5): e19428. doi:10.1371/journal.pone.0019428

8. Zhang, X.; Yazaki, J.; Sundaresan, A.; Cokus, S.; Chan, S.W.; Chen, H.; Henderson, I.R.; Shinn, P.; Pellegrini, M.; Jacobsen, S.E.; Ecker, J.R. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126:1189-201

9. Karlin, S.; Doerfler, W.; Cardon, L.R. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? Journal of Virology 68:2889-97

10. Cheng, X.; Virk, N.; Chen, W.; Ji, S.; Ji, S.; Sun, Y.; Wu, X. 2013. CpG usage in RNA viruses: data and hypotheses. PLoS One 8(9). doi: 10.1371/journal.pone.0074109

11. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. 2015. NCBI viral genomes resource. Nucleic Acids Research 43(Database issue):D571-7. doi: 10.1093/nar/gku1207. Epub 2014 Nov 26

12. R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

13. Husson, F.; Josse, J.; Le, S.; Mazet, J. 2015. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.4. URL http://CRAN.R-project.org/package=FactoMineR

14. Krupovic, M.; Ghabrial, S.A.; Jiang, D. 2015. Create one genus and one family for classification of Sclerotinia sclerotiorum hypovirulence associated DNA virus 1. International Committee on Taxonomy of Viruses: Files and Discussions

15. Elzanowski, A. & Ostell, J. 2013. The Genetic Codes. National Center for Biotechnology Information. URL http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

16. Uldahl, K. & Peng, X. 2013. Thermophilic Microbes in Environmental and Industrial Biotechnology. Biology, Biodiversity and Application of Thermophilic Viruses. pp 271-304

17. Gorbalenya, A.E.; Enjuanes, L.; Ziebuhr, J.; Snijder, E.J. Nidovirales: evolving the largest RNA virus genome. 2006. Virus Research 117(1):17-37. Epub 2006 Feb 28

18. Kühn, F.J. & Knopf, C.W. 1996. Herpes simplex virus type 1 DNA polymerase. Mutational analysis of the 3'-5'-exonuclease domain. The Journal of Biological Chemistry 271(46):29245-54

ANEXO

Cuadro A1. Carpetas del *Viral Genomes* donde faltaron los archivos .fna.

Carpetas sin archivo .fna
Acinetobacter_phage_AP205_uid14710
Caulobacter_phage_phiCb5_uid181078
Enterobacteria_phage_BZ13_uid14635
Enterobacteria_phage_C_1_INW_2012_uid184162
Enterobacteria_phage_FI_sensu_lato_uid15459
Enterobacteria_phage_Hgal1_uid184161
Enterobacteria_phage_MS2_uid14659
Enterobacteria_phage_M_uid183161
Enterobacteriophage_Qbeta_uid15479
Marine_RNA_virus_JP_A_uid20649
Marine_RNA_virus_JP_B_uid20651
Marine_RNA_virus_SOG_uid20647
Pseudomonas_phage_phi12_uid14855
Pseudomonas_phage_phi13_uid14854
Pseudomonas_phage_phi2954_uid34533
Pseudomonas_phage_phi6_uid14788
Pseudomonas_phage_phi8_uid14731
Pseudomonas_phage_PP7_uid15076
Pseudomonas_phage_PRR1_uid17481

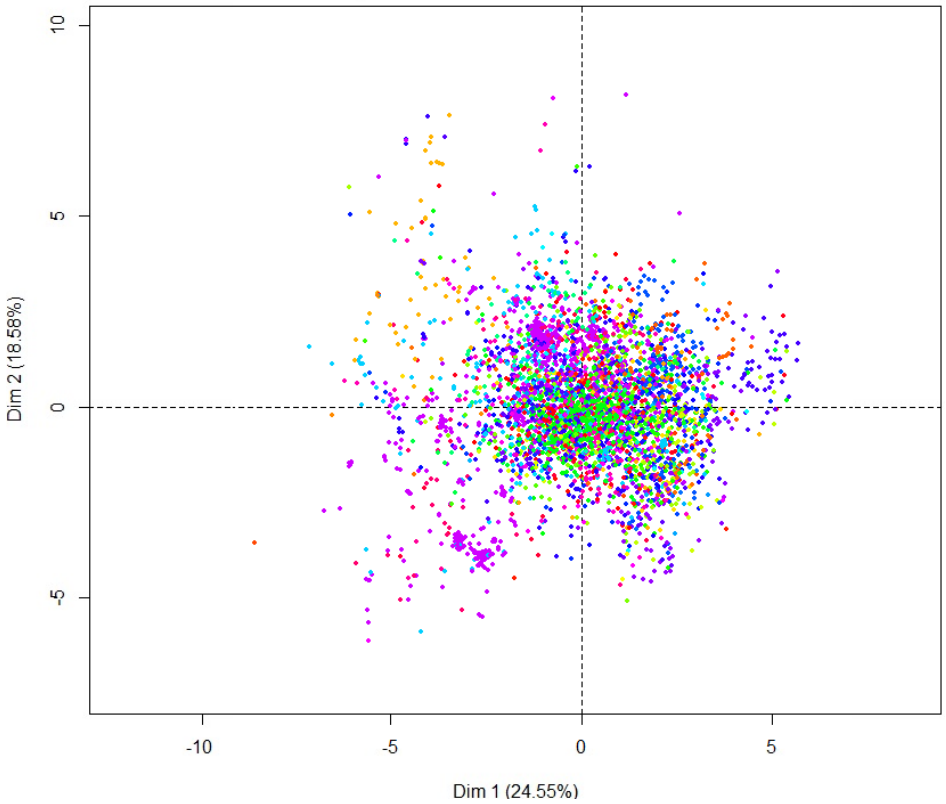

Figura A1. Dimensión 1 y 2 del análisis de componentes principales para las frecuencias relativas de dinucleótidos, coloreados por la familia a la que pertenecen.

Cuadro A2. **Orden**, **Grupo de Baltimore**, **Hospedero** y **Código genético** para cada familia viral; en rojo se señalan las familias con miembros que utilizan diferentes códigos genéticos; 1: código genético estándar; 4: código genético con UGA como triptófano; 11: código genético universal, con algunos codones de inicio alternativos además de ATG.

Familia	Orden	Grupo	Hospedero	Código genético
Adenoviridae	sin asignar	dsDNA	vertebrados	1
Alloherpesviridae	*Herpesvirales*	dsDNA	vertebrados	1
Alphaflexiviridae	*Tymovirales*	ssRNA(+)	plantas, hongos	1
Alphatetraviridae	sin asignar	ssRNA(+)	invertebrados	1
Alvernaviridae	sin asignar	ssRNA(+)	algas	1
Amalgaviridae	sin asignar	dsRNA	plantas	1
Ampullaviridae	sin asignar	dsDNA	arqueas	11
Anelloviridae	sin asignar	ssDNA	vertebrados	1
Arenaviridae	sin asignar	ssRNA(-)	vertebrados	1
Arteriviridae	*Nidovirales*	ssRNA(+)	vertebrados	1
Ascoviridae	sin asignar	dsDNA	invertebrados	1
Asfarviridae	sin asignar	dsDNA	animales	1
Astroviridae	sin asignar	ssRNA(+)	vertebrados	1
Baculoviridae	sin asignar	dsDNA	invertebrados	1
Barnaviridae	sin asignar	ssRNA(+)	hongos	1
Benyviridae	sin asignar	ssRNA(+)	plantas	1
Betaflexiviridae	*Tymovirales*	ssRNA(+)	plantas	1
Bicaudaviridae	sin asignar	dsDNA	arqueas	11
Bidnaviridae	sin asignar	ssDNA	invertebrados	1
Birnaviridae	sin asignar	dsRNA	animales	1
Bornaviridae	*Mononegavirales*	ssRNA(-)	vertebrados	1
Bromoviridae	sin asignar	ssRNA(+)	plantas	1
Bunyaviridae	sin asignar	ssRNA(-)	plantas, animales	1
Caliciviridae	sin asignar	ssRNA(+)	vertebrados	1
Carmotetraviridae	sin asignar	ssRNA(+)	invertebrados	1
Caulimoviridae	sin asignar	dsDNA(RT)	plantas	1
Chrysoviridae	sin asignar	dsRNA	hongos	1
Circoviridae	sin asignar	ssDNA	vertebrados	1
Closteroviridae	sin asignar	ssRNA(+)	plantas	1
Coronaviridae	*Nidovirales*	ssRNA(+)	vertebrados	1
Corticoviridae	sin asignar	dsDNA	bacterias	11
Cystoviridae	sin asignar	dsRNA	bacterias	11
Dicistroviridae	*Picornavirales*	ssRNA(+)	invertebrados	1
Endornaviridae	sin asignar	dsRNA	plantas, hongos	1
Filoviridae	*Mononegavirales*	ssRNA(-)	vertebrados	1
Flaviviridae	sin asignar	ssRNA(+)	animales	1
Fuselloviridae	sin asignar	dsDNA	arqueas	11
Gammaflexiviridae	*Tymovirales*	ssRNA(+)	hongos	1
Geminiviridae	sin asignar	ssDNA	plantas	1
Globuloviridae	sin asignar	dsDNA	arqueas	11
Hepadnaviridae	sin asignar	dsDNA(RT)	vertebrados	1
Hepeviridae	sin asignar	ssRNA(+)	vertebrados	1
Herpesviridae	*Herpesvirales*	dsDNA	vertebrados	1
Hypoviridae	sin asignar	dsRNA	hongos	1
Hytrosaviridae	sin asignar	dsDNA	invertebrados	1
Iflaviridae	*Picornavirales*	ssRNA(+)	invertebrados	1

Familia	Orden	Grupo	Hospedero	Código genético
Inoviridae	sin asignar	ssDNA	bacterias	4;11
Iridoviridae	sin asignar	dsDNA	animales	1
Leviviridae	sin asignar	ssRNA(+)	bacterias	11
Lipothrixviridae	*Ligamenvirales*	dsDNA	arqueas	11
Luteoviridae	sin asignar	ssRNA(+)	plantas	1
Malacoherpesviridae	*Herpesvirales*	dsDNA	invertebrados	1
Marnaviridae	*Picornavirales*	ssRNA(+)	algas	1
Marseilleviridae	sin asignar	dsDNA	protozoarios	1
Megabirnaviridae	sin asignar	dsRNA	hongos	1
Mesoniviridae	*Nidovirales*	ssRNA(+)	invertebrados	1
Microviridae	sin asignar	ssDNA	bacterias	4;11
Mimiviridae	sin asignar	dsDNA	protozoarios	1
Mycodnaviridae	sin asignar	ssDNA	animales, hongos, *environment*	1
Myoviridae	*Caudovirales*	dsDNA	procariotas	11
Nanoviridae	sin asignar	ssDNA	plantas	1
Narnaviridae	sin asignar	ssRNA(+)	hongos	1;4
Nimaviridae	sin asignar	dsDNA	invertebrados	1
Nodaviridae	sin asignar	ssRNA(+)	animales	1
Nudiviridae	sin asignar	dsDNA	animales	1
Nyamiviridae	*Mononegavirales*	ssRNA(-)	animales	1
Ophioviridae	sin asignar	ssRNA(-)	plantas	1
Orthomyxoviridae	sin asignar	ssRNA(-)	animales	1
Papillomaviridae	sin asignar	dsDNA	vertebrados	1
Paramyxoviridae	*Mononegavirales*	ssRNA(-)	vertebrados	1
Partitiviridae	sin asignar	dsRNA	plantas, hongos	1
Parvoviridae	sin asignar	ssDNA	animales	1
Permutotetraviridae	sin asignar	ssRNA(+)	invertebrados	1
Phycodnaviridae	sin asignar	dsDNA	algas	1
Picobirnaviridae	sin asignar	dsRNA	vertebrados	1
Picornaviridae	*Picornavirales*	ssRNA(+)	vertebrados	1
Plasmaviridae	sin asignar	dsDNA	bacterias	1
Podoviridae	*Caudovirales*	dsDNA	bacterias	4;11
Polydnaviridae	sin asignar	dsDNA	invertebrados	1
Polyomaviridae	sin asignar	dsDNA	vertebrados	1
Potyviridae	sin asignar	ssRNA(+)	plantas	1
Poxviridae	sin asignar	dsDNA	animales	1
Quadriviridae	sin asignar	dsRNA	hongos	1
Reoviridae	sin asignar	dsRNA	plantas, animales, hongos, algas	1
Retroviridae	sin asignar	ssRNA(RT)	vertebrados	1
Rhabdoviridae	*Mononegavirales*	ssRNA(-)	plantas, animales	1
Roniviridae	*Nidovirales*	ssRNA(+)	invertebrados	1
Rudiviridae	*Ligamenvirales*	dsDNA	arqueas	11
Secoviridae	*Picornavirales*	ssRNA(+)	plantas	1
Siphoviridae	*Caudovirales*	dsDNA	procariotas	11
Sphaerolipoviridae	sin asignar	dsDNA	procariotas	11
Tectiviridae	sin asignar	dsDNA	bacterias	11
Togaviridae	sin asignar	ssRNA(+)	animales	1
Tombusviridae	sin asignar	ssRNA(+)	plantas	1
Totiviridae	sin asignar	dsRNA	animales, hongos, protozoarios	1
Turriviridae	sin asignar	dsDNA	arqueas	11
Tymoviridae	*Tymovirales*	ssRNA(+)	plantas, invertebrados	1
Virgaviridae	sin asignar	ssRNA(+)	plantas	1

Cuadro A3. Número de virus (**n**) y las medias de las frecuencias relativas de dinucleótidos (**ρ*NN**) para cada familia; en rojo se señalan los valores menores o iguales a 0,78 y en verde los mayores o iguales a 1,23.

Familia	n	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
Adenoviridae	51	1,15	0,95	1,01	0,88	1,07	1,02	0,83	1,05	1,01	1,08	1,04	0,89	0,75	0,98	1,09	1,18
Alloherpesviridae	7	1,05	1,05	0,99	0,90	1,16	0,96	0,90	1,00	1,06	0,95	0,97	1,04	0,70	1,05	1,16	1,05
Alphaflexiviridae	45	1,06	1,01	1,05	0,86	1,13	0,96	0,73	1,15	1,11	0,99	1,05	0,82	0,66	1,05	1,22	1,15
Alphatetraviridae	3	1,02	1,12	0,96	0,85	1,06	0,86	1,16	0,94	1,15	1,02	0,83	1,01	0,70	1,06	1,01	1,21
Alvernaviridae	1	1,07	0,94	0,98	1,04	1,01	1,08	0,93	0,98	1,10	1,02	1,05	0,86	0,83	0,94	1,04	1,14
Amalgaviridae	3	0,91	0,89	1,12	1,04	1,14	1,05	0,68	1,18	1,17	0,97	1,10	0,73	0,78	1,12	0,99	1,13
Ampullaviridae	1	1,06	1,02	1,13	0,86	1,01	1,14	0,60	1,13	1,12	0,79	1,13	0,93	0,88	1,02	1,01	1,10
Anelloviridae	46	1,07	1,02	1,02	0,86	1,03	1,14	0,65	1,13	0,98	0,94	1,32	0,80	0,91	0,86	1,06	1,24
Arenaviridae	34	1,10	0,90	1,13	0,90	1,31	1,06	0,27	1,12	1,13	0,84	1,12	0,92	0,60	1,15	1,33	1,06
Arteriviridae	9	1,09	1,04	0,98	0,91	1,25	1,02	0,71	1,05	0,99	1,07	0,98	0,96	0,68	0,90	1,34	1,06
Ascoviridae	4	1,06	1,16	0,77	0,99	1,04	0,81	1,42	0,76	1,11	0,89	0,80	1,16	0,82	1,11	1,05	1,06
Asfarviridae	1	1,19	0,82	0,89	0,99	1,04	1,20	0,92	0,89	0,86	1,24	1,20	0,86	0,88	0,90	1,04	1,16
Astroviridae	36	1,06	1,01	0,98	0,94	1,27	1,14	0,51	1,05	1,02	0,99	1,08	0,92	0,68	0,89	1,35	1,10
Baculoviridae	68	1,22	1,09	0,68	0,95	1,12	0,80	1,49	0,68	0,92	1,19	0,80	1,09	0,76	0,92	1,12	1,22
Barnaviridae	1	1,13	1,01	0,91	0,96	1,03	0,95	0,94	1,06	1,12	0,81	1,02	1,01	0,74	1,20	1,10	0,99
Benyviridae	3	1,13	1,05	0,90	0,95	1,10	1,12	0,84	0,98	1,01	0,96	0,99	1,02	0,83	0,94	1,16	1,04
Betaflexiviridae	70	1,09	0,87	1,08	0,92	1,22	1,03	0,57	1,12	1,10	1,08	1,02	0,81	0,66	1,04	1,20	1,16
Bicaudaviridae	3	1,10	0,91	1,07	0,91	1,02	1,04	0,79	1,08	0,99	1,25	1,00	0,88	0,89	0,93	1,05	1,12
Bidnaviridae	1	0,96	0,83	1,13	1,06	1,04	1,18	0,62	1,05	0,99	1,14	1,21	0,86	1,03	1,04	0,93	0,98
Birnaviridae	8	0,95	1,03	1,08	0,91	1,15	1,04	0,63	1,21	1,13	0,82	1,12	0,91	0,67	1,13	1,28	0,94
Bornaviridae	3	0,94	0,95	1,11	1,01	1,23	1,04	0,56	1,10	1,13	0,91	1,10	0,85	0,77	1,09	1,16	1,02
Bromoviridae	34	1,10	0,94	0,98	0,96	0,97	1,16	0,93	0,98	1,25	0,87	0,88	0,95	0,72	1,06	1,16	1,09
Bunyaviridae	42	1,00	0,88	1,13	0,98	1,31	1,03	0,29	1,15	1,09	1,06	1,04	0,83	0,74	1,07	1,28	1,04
Caliciviridae	24	1,10	1,07	0,88	0,94	1,30	1,02	0,65	1,03	1,00	0,94	1,10	0,96	0,57	0,95	1,42	1,07
Carmotetraviridae	1	1,04	1,03	1,03	0,90	1,13	1,08	0,80	1,00	1,01	0,94	1,02	1,02	0,79	0,94	1,17	1,10
Caulimoviridae	58	1,02	0,84	1,18	0,95	1,13	1,14	0,50	1,16	1,15	0,95	1,03	0,79	0,73	1,18	1,11	1,12
Chrysoviridae	5	0,96	1,15	0,95	0,97	1,22	0,85	0,86	1,05	1,02	1,09	0,90	1,00	0,82	0,83	1,32	0,97
Circoviridae	60	1,12	0,95	0,96	0,97	0,99	1,14	0,93	0,96	1,03	0,96	1,04	0,97	0,85	0,97	1,06	1,10
Closteroviridae	35	1,12	1,03	1,00	0,87	1,03	0,98	0,93	1,02	1,17	0,84	0,94	0,98	0,75	1,11	1,07	1,12
Coronaviridae	50	1,05	1,13	0,98	0,90	1,26	1,03	0,52	1,08	0,88	1,19	0,92	1,05	0,89	0,77	1,33	1,00
Corticoviridae	1	1,27	0,88	0,85	0,89	1,10	0,88	1,21	0,76	0,76	1,62	0,96	0,90	0,83	0,65	1,07	1,38
Cystoviridae	5	1,23	0,99	0,81	1,00	1,01	0,83	1,17	0,99	1,09	1,01	0,90	1,01	0,69	1,19	1,10	0,99
Dicistroviridae	18	1,05	0,98	0,98	0,97	1,07	1,15	0,77	1,00	1,14	0,97	0,95	0,91	0,81	0,95	1,19	1,09
Endornaviridae	17	0,98	1,09	0,90	1,04	1,30	1,05	0,64	0,87	1,01	1,02	0,99	0,97	0,79	0,83	1,39	1,07
Filoviridae	8	1,02	0,92	1,06	0,99	1,19	1,09	0,53	1,04	1,10	0,93	1,17	0,82	0,74	1,07	1,18	1,12
Flaviviridae	78	1,04	1,04	0,98	0,94	1,24	1,10	0,57	1,15	1,07	0,93	1,07	0,90	0,63	0,96	1,35	1,04
Fuselloviridae	10	1,05	1,01	1,05	0,90	1,00	1,05	0,83	1,09	0,94	0,94	1,25	0,93	0,99	0,99	0,89	1,10
Gammaflexiviridae	1	1,13	1,02	0,87	0,95	1,06	1,01	0,85	1,06	1,08	1,17	0,90	0,75	0,69	0,80	1,47	1,20
Geminiviridae	368	1,11	0,87	0,92	1,05	1,00	1,25	0,82	0,97	1,11	0,91	1,03	0,95	0,82	1,01	1,17	1,02
Globuloviridae	2	1,04	1,04	0,85	1,09	0,93	1,10	0,94	1,03	0,79	1,10	1,20	0,90	1,23	0,78	0,98	1,00
Hepadnaviridae	12	1,16	0,92	0,96	0,95	1,08	1,14	0,55	1,16	1,05	0,93	1,37	0,72	0,73	1,01	1,16	1,12
Hepeviridae	4	0,99	1,02	1,01	1,00	1,21	1,07	0,72	1,00	0,98	1,04	1,02	0,96	0,81	0,86	1,28	1,04
Herpesviridae	57	1,11	0,97	0,97	0,96	1,04	1,03	0,94	0,97	1,02	1,01	1,02	0,96	0,82	1,01	1,04	1,12
Hypoviridae	10	1,10	1,03	0,93	0,95	1,16	1,00	0,79	1,04	1,13	0,89	1,04	0,93	0,62	1,08	1,22	1,09
Hytrosaviridae	2	1,04	1,00	0,72	1,10	1,18	1,04	1,23	0,69	1,04	1,00	0,96	1,03	0,81	1,04	1,19	1,04
Iflaviridae	19	1,00	0,96	0,97	1,04	1,05	1,23	0,83	0,94	1,04	1,07	0,96	0,96	0,94	0,87	1,15	1,03

Familia	n	ρ*AA	ρ*AC	ρ*AG	ρ*AT	ρ*CA	ρ*CC	ρ*CG	ρ*CT	ρ*GA	ρ*GC	ρ*GG	ρ*GT	ρ*TA	ρ*TC	ρ*TG	ρ*TT
Inoviridae	37	1,22	0,94	0,83	1,04	1,08	0,94	1,06	0,96	0,99	1,20	1,00	0,90	0,72	0,98	1,14	1,08
Iridoviridae	16	1,13	0,98	0,96	0,88	1,06	1,18	0,80	0,97	1,00	0,94	1,14	0,98	0,80	0,99	1,08	1,14
Leviviridae	11	1,19	0,99	0,84	0,96	0,97	0,95	1,08	1,00	1,05	0,98	1,05	0,94	0,82	1,07	1,03	1,08
Lipothrixviridae	8	1,02	1,03	1,05	0,94	1,10	0,98	0,87	1,00	1,03	0,87	1,01	1,04	0,91	1,06	1,02	1,04
Luteoviridae	29	1,14	0,90	1,04	0,91	1,04	1,07	0,81	1,07	1,12	0,93	1,05	0,88	0,68	1,11	1,10	1,16
Malacoherpesviridae	2	1,14	0,95	0,95	0,93	1,09	1,04	0,91	0,93	1,15	0,86	1,04	0,94	0,68	1,15	1,08	1,15
Marnaviridae	1	1,18	1,00	0,90	0,93	1,14	1,04	0,95	0,88	1,05	1,09	0,91	0,97	0,68	0,89	1,22	1,18
Marseilleviridae	3	1,35	0,82	1,09	0,73	0,98	1,02	0,92	1,07	1,35	0,79	1,00	0,82	0,39	1,33	0,97	1,36
Megabirnaviridae	1	1,19	1,09	0,78	1,02	1,15	0,84	0,94	1,09	0,97	1,24	0,92	0,89	0,74	0,81	1,32	1,02
Mesoniviridae	6	0,96	1,15	0,85	1,02	1,14	0,96	0,83	0,95	0,97	1,13	0,98	0,95	0,96	0,78	1,33	1,04
Microviridae	19	1,19	0,91	0,91	0,94	0,95	0,92	0,98	1,12	1,05	1,16	1,00	0,86	0,81	1,04	1,11	1,07
Mimiviridae	5	1,05	0,85	0,78	1,08	1,20	1,32	0,69	0,80	1,07	0,90	1,35	0,84	0,86	1,08	1,20	1,05
Mycodnaviridae	23	1,06	0,95	0,91	1,08	1,06	1,04	0,98	0,92	1,13	0,87	1,02	1,00	0,73	1,15	1,09	1,01
Myoviridae	300	1,07	0,98	0,96	1,00	1,09	0,94	0,94	0,99	1,06	1,06	0,95	0,95	0,77	1,04	1,10	1,08
Nanoviridae	8	1,01	0,86	0,98	1,09	1,00	1,21	0,92	0,94	1,07	1,01	0,97	0,94	0,94	1,01	1,09	0,99
Narnaviridae	13	1,03	0,95	1,02	1,00	0,97	1,17	0,67	1,11	1,17	0,78	1,18	0,87	0,88	1,12	1,04	1,01
Nimaviridae	1	1,11	0,91	1,05	0,91	1,11	1,11	0,64	1,06	1,12	0,86	1,11	0,90	0,72	1,11	1,13	1,12
Nodaviridae	15	1,06	1,08	0,90	0,97	1,16	0,95	0,95	0,94	1,05	0,99	0,91	1,05	0,71	0,99	1,25	1,05
Nudiviridae	4	1,06	1,03	0,84	1,01	1,03	0,91	1,36	0,85	1,03	0,99	0,90	1,03	0,89	1,03	1,04	1,06
Nyamiviridae	4	1,15	0,75	1,34	0,81	1,18	1,04	0,56	1,22	1,17	0,96	1,04	0,86	0,56	1,17	1,15	1,06
Ophioviridae	4	1,09	0,80	1,19	0,97	1,29	0,98	0,41	1,03	1,23	0,88	1,01	0,90	0,67	1,21	1,20	1,05
Orthomyxoviridae	11	1,03	0,92	1,04	0,98	1,23	1,10	0,41	1,19	1,15	0,95	1,07	0,76	0,61	1,10	1,35	1,12
Papillomaviridae	128	1,01	1,03	1,05	0,93	1,21	1,18	0,51	1,06	0,97	1,07	1,08	0,91	0,85	0,78	1,24	1,11
Paramyxoviridae	60	0,94	0,91	1,10	1,06	1,23	1,07	0,44	1,10	1,11	0,93	1,11	0,84	0,80	1,11	1,22	0,98
Partitiviridae	37	1,19	1,00	0,84	0,92	0,97	0,96	0,99	1,08	1,21	0,95	0,91	0,90	0,72	1,09	1,19	1,06
Parvoviridae	84	1,04	1,06	0,97	0,93	1,14	0,98	0,71	1,09	1,04	0,97	1,11	0,88	0,79	0,97	1,19	1,11
Permutotetraviridae	1	1,12	0,91	1,05	0,90	0,91	1,33	0,80	1,00	1,15	0,90	1,03	0,92	0,78	0,88	1,12	1,21
Phycodnaviridae	16	1,10	0,99	0,84	1,02	1,11	1,01	1,06	0,85	1,12	0,86	1,03	0,98	0,73	1,12	1,10	1,10
Picobirnaviridae	1	1,02	1,02	0,90	1,06	1,18	0,95	0,81	1,03	0,99	1,04	0,99	0,99	0,85	0,98	1,27	0,92
Picornaviridae	83	1,11	1,00	0,97	0,92	1,20	1,15	0,49	1,09	1,10	0,93	1,11	0,89	0,66	0,93	1,36	1,07
Plasmaviridae	1	1,02	1,04	0,94	1,00	1,18	0,98	0,59	1,03	0,96	1,08	1,15	0,93	0,92	0,93	1,18	1,03
Podoviridae	219	1,05	1,01	1,00	0,95	1,10	0,93	0,87	1,08	1,03	1,07	0,95	0,98	0,80	1,01	1,16	1,01
Polydnaviridae	5	1,09	0,99	0,90	0,98	1,11	0,92	1,08	0,90	1,01	1,08	0,92	0,99	0,84	1,02	1,10	1,09
Polyomaviridae	65	1,13	0,82	1,21	0,85	1,19	1,22	0,20	1,21	0,91	1,13	1,29	0,79	0,79	0,93	1,16	1,14
Potyviridae	123	1,02	1,02	0,99	0,97	1,36	0,81	0,71	0,95	1,08	1,11	0,93	0,88	0,64	1,02	1,29	1,17
Poxviridae	37	0,99	0,97	0,91	1,04	1,02	1,06	1,11	0,91	1,07	0,94	1,07	0,97	0,94	1,07	1,02	0,99
Quadriviridae	1	0,82	1,20	1,00	1,00	1,38	0,65	0,85	1,13	1,05	1,16	0,84	0,96	0,72	0,91	1,42	0,90
Reoviridae	58	0,97	1,02	0,93	1,07	1,13	0,91	0,97	0,93	1,16	1,03	0,89	0,90	0,82	1,02	1,16	1,05
Retroviridae	63	1,08	0,89	1,11	0,92	1,05	1,23	0,51	1,17	1,03	0,93	1,25	0,82	0,83	0,96	1,14	1,09
Rhabdoviridae	80	0,98	0,89	1,06	1,05	1,20	1,12	0,48	1,11	1,24	0,75	1,14	0,79	0,67	1,26	1,21	1,02
Roniviridae	1	0,90	1,05	0,89	1,13	1,36	0,74	0,87	0,98	1,02	0,95	1,08	0,99	0,68	1,29	1,22	0,88
Rudiviridae	5	1,10	0,89	0,96	0,95	1,05	1,01	0,95	0,94	1,08	1,16	1,00	0,87	0,84	1,06	1,03	1,14
Secoviridae	41	1,14	0,88	1,03	0,93	1,21	1,12	0,55	1,08	1,08	1,09	1,05	0,83	0,67	0,95	1,24	1,14
Siphoviridae	678	1,06	1,01	0,92	1,01	1,07	0,89	1,04	0,97	1,09	1,09	0,93	0,96	0,70	1,07	1,08	1,06
Sphaerolipoviridae	4	1,02	1,08	1,00	0,81	0,93	1,03	0,98	1,06	1,17	0,85	1,05	1,02	0,79	1,18	0,90	1,02
Tectiviridae	5	1,19	1,01	0,79	0,98	1,01	1,05	1,11	0,84	0,92	1,07	1,12	0,91	0,86	0,86	1,07	1,23
Togaviridae	24	0,97	1,07	1,00	0,95	1,13	0,98	0,88	1,00	1,03	1,04	0,91	1,02	0,84	0,90	1,24	1,04
Tombusviridae	57	1,04	1,03	0,99	0,94	1,17	1,07	0,74	1,04	1,01	0,91	1,08	0,98	0,79	0,99	1,17	1,04
Totiviridae	43	1,02	1,12	0,90	0,96	1,08	0,92	0,95	1,03	1,05	0,99	0,99	0,98	0,84	0,99	1,13	1,03
Turriviridae	2	1,01	0,85	1,19	0,93	0,97	1,00	0,83	1,17	1,11	1,13	1,02	0,75	0,92	1,10	0,82	1,20
Tymoviridae	30	1,26	0,86	1,16	0,87	1,12	0,94	0,79	1,12	1,26	0,92	1,12	0,84	0,41	1,26	1,10	1,03
Virgaviridae	47	1,11	1,04	1,00	0,87	1,06	0,95	0,91	1,04	1,13	0,95	0,90	0,99	0,74	1,04	1,13	1,11